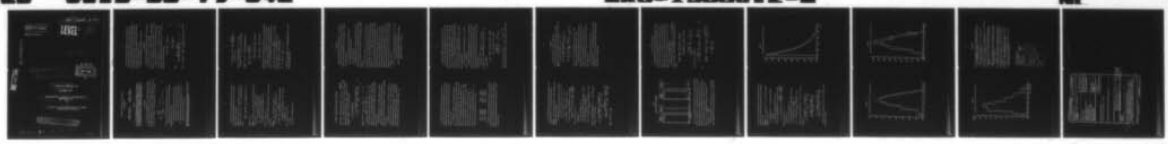
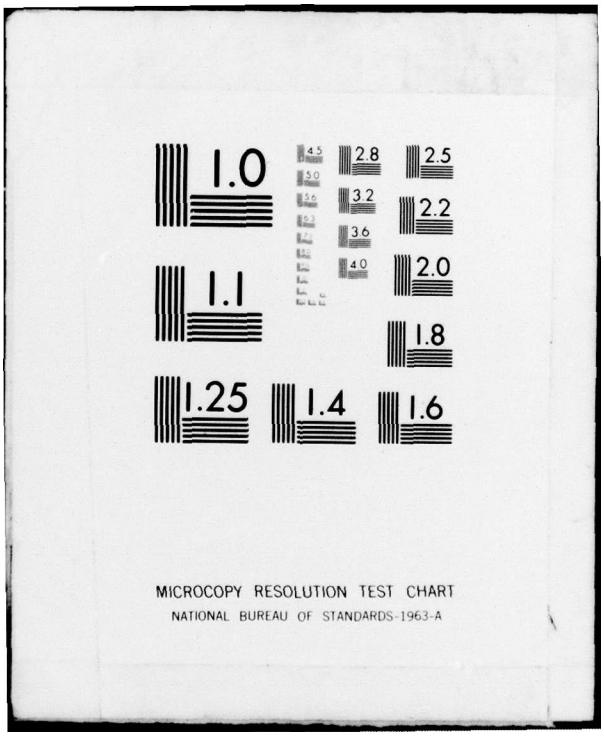
AD-A080 601 WISCONSIN UNIV-MADISON DEPT OF STATISTICS F/G 12/1
TRANSFORMATION OF GROUPED OR CENSORED DATA TO NEAR NORMALITY.(U)
NOV 79 V M GUERRERO, R A JOHNSON N00014-78-C-0722
UNCLASSIFIED UWIS-DS-79-542 ARO-15660.2-M NL

1 OF 1
AD A080601



END
DATE FILMED
3-80
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN

Madison, Wisconsin

(14) UWIS-DS-79-542

(15) N0014-78-C-0722

(9) TECHNICAL REPORT, NO. 542

(11) November 1979

(6) TRANSFORMATION OF GROUPED OR CENSORED DATA TO NEAR NORMALITY

by

(10) Victor M. Guerrero and Richard A. Johnson
University of Wisconsin

Transformation of Grouped or Censored Data to Near Normality

by

Victor M. Guerrero
University of Wisconsin

Richard A. Johnson
University of Wisconsin

Summary

Box and Cox (1964) proposed a power transformation which has proven utility for transforming ungrouped data to near normality. In this paper, we extend its applicability to two frequently encountered situations: (i) grouped data and (ii) Type I censored data. The asymptotic properties of the estimators are derived and illustrative examples are presented.

1. Introduction

Let us consider a random sample $X_1,\ldots X_n$ from an absolutely continuous distribution with probability density function (pdf) g concentrated on $(0,\infty)$. We assume the existence of a unique vector of parameter values $\underset{\sim}{\theta}_0 = (\mu_0,\sigma_0,\lambda_0)'$ such that the distribution of the transformed variables $X_j^{(\lambda_0)}$ defined by

$$X_j^{(\lambda_0)} = \begin{cases} \dfrac{X_j^{\lambda_0}-1}{\lambda_0} & \text{if} \quad \lambda_0 \neq 0 \\ \log(X_j) & \text{if} \quad \lambda_0 = 0 \end{cases} \qquad \text{for} \quad j=1,\ldots,n \tag{1}$$

is "closest" to a normal distribution with parameters μ_0 and σ_0^2.

The problem to be considered here is that of obtaining the Maximum Likelihood Estimator (MLE) of $\underset{\sim}{\theta}$ when:

(i) the original random variables are unobserved and the only available information is the number of observations falling within arbitrary, but specified, intervals of the real line, or

(ii) some exact observations are available besides the counts of observations in other specified intervals.

*This research was sponsored by the Office of Naval Research under Grant No. N00014-78-C-0722.(Also funded by Army Research Office).

Usually, when dealing with incomplete data, explicit expressions for the MLE's are not available. However, it is possible to gain some insight into the proposed procedure by studying their asymptotic properties as in Hernández and Johnson (1979). Thus we establish strong consistency and asymptotic normality of the MLE and identify its limit with a minimum Kullback-Leibler information number property.

Examples of both grouped and censored data are considered.

2. Grouped Data

Let the sample be grouped into $k(\geq 3)$ prespecified intervals denoted by $D_1 = [a_0,a_1)$, $D_2 = [a_1,a_2),\ldots,D_k = [a_{k-1},a_k)$ where $0 = a_0 < a_1 <\ldots< a_{k-1} < a_k = \infty$. The count of the number of observations in interval D_i will be denoted by n_i, and the total sample size is $n = \sum_{i=1}^{k} n_i$.

In order to obtain the MLE of $\underset{\sim}{\theta}$, we tentatively assume (pretend), that $X_j^{(\lambda_0)} \sim N(\mu_0,\sigma_0^2)$ for $j = 1,\ldots,n$. Because X_j is positive, $X_j^{(\lambda_0)}$ cannot have a normal distribution, except possibly for $\lambda_0 = 0$. Under this assumption the log-likelihood becomes

$$\ell_n(\underset{\sim}{\theta}) = \log(n!) - \sum_{i=1}^{k} \log(n_i!) + \sum_{i=1}^{k} n_i \log[p_i(\underset{\sim}{\theta})] \tag{2}$$

where

$$p_i(\underset{\sim}{\theta}) = \Phi\left(\frac{a_i^{(\lambda)}-\mu}{\sigma}\right) - \Phi\left(\frac{a_{i-1}^{(\lambda)}-\mu}{\sigma}\right) \quad \text{for} \quad i = 1,\ldots,k \tag{3}$$

$a_0^{(\lambda)} = -\infty$ for every λ and $\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$ with $\phi(t) = (2\pi)^{-1/2} e^{-t^2/2}$.

We state without proofs two lemmas which are used to establish (1) strong consistency and (2) asymptotic normality of the MLE $\hat{\underset{\sim}{\theta}}_n$ (see Guerrero (1979) for the proofs.)

Lemma 1. Let $D_i = [a_{i-1}, a_i)$, $q_i = \int_{D_i} g(x)dx$, $\hat{p}_{i,n} = n_i/n$ and $p_i(\underset{\sim}{\theta})$ be as in (3) for $i = 1,\ldots,k$. For Ω a compact subset of $\mathbb{R}^3$ we have

$$\left| \sum_{i=1}^{k} \hat{p}_{i,n} \log[p_i(\underset{\sim}{\theta})] - \sum_{i=1}^{k} q_i \log[p_i(\underset{\sim}{\theta})] \right| \xrightarrow{a.s.} 0 \text{ as } n \to \infty$$

uniformly in $\underset{\sim}{\theta} \in \Omega$.

Lemma 2. For fixed n, let $\ell_n(\underset{\sim}{\theta})$ be the log-likelihood function (2) and $\hat{\underset{\sim}{\theta}}_n$ the MLE of $\underset{\sim}{\theta}$. For Ω a compact set, assume

(i) f is a continuous function from Ω to $\mathbb{R}$,

(ii) $\frac{1}{n} \ell_n(\underset{\sim}{\theta}) \xrightarrow{a.s.} f(\underset{\sim}{\theta})$ uniformly in $\underset{\sim}{\theta} \in \Omega$,

(iii) $f(\underset{\sim}{\theta})$ has a unique global maximum at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$.

Then, $\hat{\underset{\sim}{\theta}}_n \xrightarrow{a.s.} \underset{\sim}{\theta}_0$ as $n \to \infty$.

We now proceed to state and sketch the proof of the main result of this section.

Theorem 3. Let q_i and $p_i(\underset{\sim}{\theta})$ be as in Lemma 1. If

(i) the parameter space is a compact subset of $\mathbb{R}^3$

(ii) $H(\underset{\sim}{\theta}) = \sum_{i=1}^{k} q_i \log\left[\frac{p_i(\underset{\sim}{\theta})}{q_i}\right]$ has a unique global maximum as a function of $\underset{\sim}{\theta} = (\theta_1,\theta_2,\theta_3)' = (\mu,\sigma,\lambda)'$, and this is attained at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$.

Then, (1) $\hat{\underset{\sim}{\theta}}_n \xrightarrow{a.s.} \underset{\sim}{\theta}_0$ as $n \to \infty$.

If further,

(iii) $\underset{\sim}{\theta}_0$ is an interior point of Ω,

(iv) the Hessian of $H(\underset{\sim}{\theta})$, $\nabla^2 H(\underset{\sim}{\theta}) = \left(\frac{\partial^2 H(\underset{\sim}{\theta})}{\partial\theta_u \partial\theta_v}\right)_{3\times 3}$ is nonsingular at $\underset{\sim}{\theta}_0$.

Then, (2) $\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{\mathcal{L}} N_3(\underset{\sim}{0}, VWV')$ as $n \to \infty$, where $V = [\nabla^2 H(\underset{\sim}{\theta}_0)]^{-1}$ and $W = (w_{uv})_{3\times 3}$ is given by (4) below.

Proof: Let $\hat{p}_{i,n} = \frac{n_i}{n}$. Stirling's approximation yields

$$\left| \frac{1}{n} \ell_n(\underset{\sim}{\theta}) - \sum_{i=1}^{k} q_i \log\left[\frac{p_i(\underset{\sim}{\theta})}{q_i}\right] \right|$$

$$\leq \left| \sum_{i=1}^{k} \hat{p}_{i,n} \log[p_i(\underset{\sim}{\theta})] - \sum_{i=1}^{k} q_i \log[p_i(\underset{\sim}{\theta})] \right|$$

$$+ \left| \sum_{i=1}^{k} \hat{p}_{i,n} \log(\hat{p}_{i,n}) - \sum_{i=1}^{k} q_i \log(q_i) \right| + o(1)$$

Because of Lemma 1, the right hand side converges to zero, with probability one, uniformly in $\underset{\sim}{\theta} \in \Omega$. The consistency of $\hat{\underset{\sim}{\theta}}_n$ follows from Lemma 2.

To establish asymptotic normality, we obtain the gradient and the Hessian of the log-likelihood function by straightforward differentiation. The Hessian of $\ell_n(\underset{\sim}{\theta})$, with elements $\frac{\partial^2 \ell_n(\underset{\sim}{\theta})}{\partial\theta_u \partial\theta_v}$ for $u,v=1,2,3$ is readily seen to be continuous on Ω. So, using Taylor's formula,

$$\frac{1}{\sqrt{n}} \nabla\ell_n(\hat{\underset{\sim}{\theta}}_n) = \frac{1}{\sqrt{n}} \nabla\ell_n(\underset{\sim}{\theta}_0) + \frac{1}{n} \nabla^2 \ell_n(\underset{\sim}{\theta}_{*n})[\sqrt{n}\,(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0)],$$

with $\underset{\sim}{\theta}_{*n} = \gamma_n \underset{\sim}{\theta}_0 + (1 - \gamma_n)\hat{\underset{\sim}{\theta}}_n$, $0 < \gamma_n < 1$.

We conclude that $\frac{1}{\sqrt{n}} \nabla\ell_n(\underset{\sim}{\theta}_0)$ and $-\frac{1}{n} \nabla^2 \ell_n(\underset{\sim}{\theta}_{*n})[\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0)]$ have the same limiting distribution. Write

$$\frac{1}{n}\nabla \ell_n(\underset{\sim}{\theta}_0) = \frac{1}{n}\sum_{j=1}^{n}\left[\sum_{i=1}^{k} I_{D_i}(X_j)\underset{\sim}{a}_i(\underset{\sim}{\theta}_0)\right] = \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{Z}_j(\underset{\sim}{\theta}_0)$$

where $\underset{\sim}{a}_i(\underset{\sim}{\theta}_0) = (a_{i1}(\underset{\sim}{\theta}_0), a_{i2}(\underset{\sim}{\theta}_0), a_{i3}(\underset{\sim}{\theta}_0))'$ with $a_{ir}(\underset{\sim}{\theta}_0) = \left.\frac{\partial \log[p_i(\underset{\sim}{\theta})]}{\partial \theta_r}\right|_{\underset{\sim}{\theta}_0}$ for $r = 1,2,3$. Asymptotic normality follows from the multivariate Central Limit Theorem since the random vectors $\underset{\sim}{Z}_1(\underset{\sim}{\theta}_0), \ldots, \underset{\sim}{Z}_n(\underset{\sim}{\theta}_0)$ are iid with $E_g[\underset{\sim}{Z}_1(\underset{\sim}{\theta}_0)] = \nabla H(\underset{\sim}{\theta}_0) = \underset{\sim}{0}$ and have covariance elements

$$w_{uv} = \sum_{i=1}^{k} q_i a_{iu}(\underset{\sim}{\theta}_0) a_{iv}(\underset{\sim}{\theta}_0)$$

$$= \sum_{i=1}^{k} q_i \left(\left.\frac{\partial \log[p_i(\underset{\sim}{\theta})]}{\partial \theta_u}\right|_{\underset{\sim}{\theta}_0}\right)\left(\left.\frac{\partial \log[p_i(\underset{\sim}{\theta})]}{\partial \theta_v}\right|_{\underset{\sim}{\theta}_0}\right). \qquad (4)$$

□

<u>Remark:</u> Setting $Q = (q_i | q_i = \int_{D_i} g(x)dx > 0,\ i = 1,\ldots,k)$ and $N = (p_i(\underset{\sim}{\theta}) | p_i(\underset{\sim}{\theta}) = \Phi\left(\frac{a_i^{(\lambda)} - \mu}{\sigma}\right) - \Phi\left(\frac{a_{i-1}^{(\lambda)} - \mu}{\sigma}\right),\ i = 1,\ldots,k)$, it follows easily that, under the conditions of Theorem 3, $\underset{\sim}{\theta}_0$ is also that value of $\underset{\sim}{\theta}$ which minimizes the Kullback-Leibler information number $I[Q,N]$. A similar interpretation for ungrouped data is given by Hernaddez and Johnson (1979).

Hence, obtaining the MLE of $\underset{\sim}{\theta}$ using the (wrong) log-likelihood function (2) is asymptotically equivalent to finding the minimum of the Kullback-Leibler information number between the true probability distribution Q and a normal.

The importance of the true pdf g is reflected in the asymptotic variance of $\hat{\underset{\sim}{\theta}}_n$ derived in Theorem 3. If one is faced with the task of transforming a grouped sample to near normality, the true probabilities q_i can be estimated by the observed frequencies $\hat{p}_{i,n}$. Doing this, one obtains a consistent estimate of the asymptotic variance of $\hat{\underset{\sim}{\theta}}_n$.

3. Application

Because grouping results in loss of information, it is preferable to transform the original (ungrouped) data. Only when the original data are not available should grouped observations be transformed. We may want to transform to normality

(a) for better description or for model building purposes,

(b) to interpolate between the given data,

(c) to smooth count data.

The technique we propose will probably be most appropriate when the sample size is moderate so that the large sample theory gives some assurances, yet not so large that one would use splines or other nonparametric smoothing procedures.

In order to find the MLE of $\underset{\sim}{\theta} = (\mu,\sigma,\lambda)'$ in our transformed normal approach, we need to use an iterative procedure and this requires some initial estimates. Several methods have been devised for obtaining approximate MLE's for μ and σ (c.f. Lindley (1950), Benn and Sidebottom (1976)) which may be used for getting initial values for those parameters. But the problem still remains with respect to selecting an initial value for λ. In practice, we can apply a specialization to our case, of a two-stage procedure proposed by Richards (1961) (the assumptions under which this procedure is valid were wrongly stated by Richards, but later Kale (1963) validated the method by assuming that the conditions of the Implicit Function Theorem hold). This method enables us to obtain the value $\hat{\underset{\sim}{\theta}}_n = (\hat{\mu}_n,\hat{\sigma}_n,\hat{\lambda}_n)'$, for n fixed, as follows:

Step (1) fix λ and maximize $\ell_n(\underset{\sim}{\theta})$ with respect to μ and σ, thus obtaining a value $\hat{\underset{\sim}{\theta}}_n(\lambda) = (\hat{\mu}_n(\lambda),\hat{\sigma}_n(\lambda),\lambda)'$

Step (2) maximize the already partially maximized log-likelihood $\ell_n(\hat{\underset{\sim}{\theta}}_n(\lambda))$ with respect to λ to get $\hat{\underset{\sim}{\theta}}_n$.

In practice it is not possible to get explicit expressions for $\hat{\mu}_n(\lambda)$ and $\hat{\sigma}_n(\lambda)$, therefore the procedure relies heavily upon the use of the Implicit Function Theorem to assure the existence of such quantities. To find the values $\hat{\underset{\sim}{\theta}}_n(\lambda)$ we solve, iteratively, the maximum likelihood equations for μ and σ.

We illustrate our procedure for the life length distribution in Figure 1. The data are the grouped ages of the Mexican population in 1966 which appear in Keyfitz and Flieger (1971 p. 344). Because of the large sample size, other smoothing techniques may be more appropriate. The value of λ which maximized the log-likelihood function is $\hat{\lambda} = .34995$ and for this value we obtained $\hat{\mu} = 4.775151$ and $\hat{\sigma} = 2.712447$. These values were used to draw the histogram of the transformed data with an overlay of the appropriate normal pdf in Figure 2. Shown in Figure 1 is the corresponding transformed normal pdf. From the estimated variance-covariance matrix of $\hat{\underset{\sim}{\theta}}$ we have

$\hat{Var}(\hat{\mu}) = .000000726$ $\quad$ $\hat{Var}(\hat{\sigma}) = .000000634$

$\hat{Var}(\hat{\lambda}) = .000000011$ $\quad$ $\hat{Cov}(\hat{\mu},\hat{\sigma}) = .000000576$

$\hat{Cov}(\hat{\mu},\hat{\lambda}) = .000000079$ $\quad$ $\hat{Cov}(\hat{\sigma},\hat{\lambda}) = .000000078$.

From Figures 1 and 2 we observe that strict normality was not achieved, although the transformation did yield a nearly symmetrical distribution. Draper and Cox (1969) noticed the fact that in some cases of ungrouped data, even when the transformation procedure does not yield normality, it helps to "regularize" data.

4. Grouped Data Combined with Exact Observations in the Tails.

Here we consider a censoring scheme where exact values are recorded only for observations either in the lower or the upper tail, or for both extremes. The rest of the sample is grouped. More precisely, the values of all observations <u>below</u> the (specified) fixed value a_1 or <u>above</u> the fixed value a_{k-1} are recorded exactly, while those in between are grouped into the intervals $[a_1,a_2),\ldots,[a_{k-2},a_{k-1})$, where $0 < a_1 < \ldots < a_{k-1} < \infty$. This is a form of fixed time censoring which produces a mixed sample of exact and grouped data.

As before, we denote the intervals $[0,a_1),[a_1,a_2),\ldots,[a_{k-1},\infty)$ by $D_1,D_2,\ldots,D_k$ respectively.

The motivation behind this formulation is that it is good statistical practice to report extreme values exactly for further scrutiny. This may be done either to check for outliers or to delineate unique features of the phenomenon under study, particularly those related to the tail behavior of the distributions.

Pretending that $x_j^{(\lambda_0)} \sim N(\mu_0,\sigma_0^2)$, $j = 1,\ldots,n$ for some $\underset{\sim}{\theta}_0 = (\mu_0,\sigma_0,\lambda_0)'$, the log-likelihood of the (mixed) sample, of size $n = \sum_{i=1}^{k} n_i$, becomes

$$\ell_n(\underset{\sim}{\theta}|x) = \log(n!) - \sum_{i=1}^{k} \log(n_i!) + \sum_{i=2}^{k-1} n_i \log[p_i(\underset{\sim}{\theta})] - \frac{1}{2}(n_1 + n_k)\log(2\pi) - (n_1 + n_k)\log(\sigma) - \frac{1}{2}\sum_{j}^{n_1,n_k}\left(\frac{x_j^{(\lambda)} - \mu}{\sigma}\right)^2 + (\lambda-1)\sum_{j}^{n_1,n_k}\log(x_j). \qquad (5)$$

where $p_i(\underset{\sim}{\theta})$ is defined in (3) for $i = 2,\ldots,k-1$ and where $\sum_{j}^{n_1,n_k}$ stands for the sum over the smallest n_1 and the largest n_k order statistics.

Now, let $\ell_1^{(u)}(\underset{\sim}{\theta}|X)$ denote the log-likelihood corresponding to one ungrouped observation. Then we obtain the following theorem.

Theorem 4. Let q_i and $p_i(\underset{\sim}{\theta})$ be defined as in Lemma 1 and suppose the following conditions are satisfied:

(i) the parameter space Ω is the compact subset of R^3 defined by $\Omega = (\underset{\sim}{\theta} = (\mu,\sigma,\lambda)' \mid |\mu| \leq M,\ s_1 \leq \sigma \leq s_2,\ a \leq \lambda \leq b$ for some $0 < M, s_1, s_2, b < \infty$ and $-\infty < a < 0)$,

(ii) the moments $E_g(X^{2a})$ and $E_g(X^{2b})$ are finite,

(iii) $H(\underset{\sim}{\theta}) + E_g[I_{D_1 \cup D_k}(X)\, \ell_1^{(u)}(\underset{\sim}{\theta}|X)]$ has a unique global maximum at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$, where $H(\underset{\sim}{\theta}) = \sum_{i=2}^{k-1} q_i \log\left[\frac{p_i(\underset{\sim}{\theta})}{q_i}\right]$.

Then, (1) $\hat{\underset{\sim}{\theta}}_n \xrightarrow{a.s.} \underset{\sim}{\theta}_0$ as $n \to \infty$.

Furthermore, if:

(iv) $\underset{\sim}{\theta}_0$ is an interior point of Ω,

(v) both $E_g[X^a \log(X)]^2$ and $E_g[X^b \log(X)]^2$ are finite,

(vi) $\nabla H(\underset{\sim}{\theta}_0) + E_g[I_{D_1 \cup D_k}(X) \cdot \nabla \ell_1^{(u)}(\underset{\sim}{\theta}_0|X)] = \underset{\sim}{0}$,

(vii) $V = (\nabla^2 H(\underset{\sim}{\theta}_0) + E_g[I_{D_1 \cup D_k}(X)\, \nabla^2 \ell_1^{(u)}(\underset{\sim}{\theta}_0|X)])^{-1}$ exists.

Then, (2) $\sqrt{n}\,(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{\mathcal{L}} N_3(\underset{\sim}{0}, VWV')$ as $n \to \infty$, with the elements of $W = (w_{uv})$, $u,v = 1,2,3$, given by

$$w_{uv} = \sum_{i=2}^{k-1} q_i \left(\frac{\partial \log[p_i(\underset{\sim}{\theta})]}{\partial \theta_u}\bigg|_{\underset{\sim}{\theta}_0} \right) \left(\frac{\partial \log[p_i(\underset{\sim}{\theta})]}{\partial \theta_v}\bigg|_{\underset{\sim}{\theta}_0} \right)$$

$$+ E_g\left[I_{D_1 \cup D_k}(X) \left(\frac{\partial\, \ell_1^{(u)}(\underset{\sim}{\theta}|X)}{\partial \theta_u}\bigg|_{\underset{\sim}{\theta}_0} \right) \left(\frac{\partial\, \ell_1^{(u)}(\underset{\sim}{\theta}|X)}{\partial \theta_v}\bigg|_{\underset{\sim}{\theta}_0} \right) \right].$$

Proof: See Guerrero (1979).

5. An Example

Consider the situation in which the observations below some fixed point a_1 and above the fixed point a_{k-1} are known exactly and the rest of the sample is given in grouped form. Writing A_i for $\frac{a_i^{(\lambda)} - \mu}{\sigma}$ the likelihood equations for μ and σ are

$$\frac{\partial \ell_n}{\partial \mu} = \sum_{i=2}^{k-1} \frac{n_i}{\sigma} \left[\frac{\phi(A_{i-1}) - \phi(A_i)}{p_i(\underset{\sim}{\theta})} \right] + \frac{1}{\sigma^2} \left(\sum_{j=1}^{h} x_j^{(\lambda)} - h\mu \right)$$

$$\frac{\partial \ell_n}{\partial \sigma} = \sum_{i=2}^{k-1} \frac{n_i}{\sigma} \left[\frac{A_{i-1}\phi(A_{i-1}) - A_i\phi(A_i)}{p_i(\underset{\sim}{\theta})} \right] - \frac{h}{\sigma} + \frac{1}{\sigma^3} \sum_{j=1}^{h} (x_j^{(\lambda)} - \mu)^2$$

with $n = n_1 + n_k$ the total number of exact observations. These equations are to be solved iteratively for fixed λ values.

Data approximating our assumptions come from the Wage Surveys, in which a category like "under 100" has a footnote giving counts for such a category in much smaller intervals. For the purpose of illustration we use the data on Weekly Earnings of Secretaries (in Hartford, CT) given in Table 1 below.

We will assume that observations outside the range 130-250 are given exactly so the five observations falling in [120, 130) are placed at 121, 123, 125, 127, and 129, and the last three observations will be taken as 255, 285 and 325.

Table 1

WEEKLY EARNINGS OF SECRETARIES

Weekly Earnings of	No. of Secretaries	Weekly Earnings of	No. of Secretaries
120 and under 130	5	230 and under 240	24
130 140	45	240 250	4
140 150	125	250 260	1
150 160	126	260 270	-
160 170	141	270 280	-
170 180	100	280 290	1
180 190	69	290 300	-
190 200	70	300 310	-
200 210	48	310 320	-
210 220	43	320 330	1
220 230	11	Total	814

Source: Industry Wage Surveys, December 1976. Bulletin 1988 (Table 75).

A plot of the transformed histogram, with the "closest" normal approximation is shown as Figure 3 below. It can be observed that even though exact normality has not been achieved, the transformation worked reasonably well. The improvement in normality can also be appreciated by fitting the inverted normal pdf to the original data as shown in Figure 4. The MLE's in this case were found to be $\hat{\lambda} = -.99$ (approximately the reciprocal transformation), $\hat{\mu} = 1.003809$ and $\hat{\sigma} = .000891$.

The chi-square statistic for a goodness-of-fit test is $X^2 = 63.59$ which rejects normality because of the drops in intervals [180, 190) and [220, 230).

6. Type I Censoring Procedures

Consider data that occurs when a censoring scheme operates in such a way that observations falling between the specified fixed limits a_1 and a_2 are recorded exactly, while only the counts of values below a_1 and above $a_2 (0 < a_1 < a_2 < \infty)$ are available. This is the usual double Type I censoring scheme which, for our case, divides the positive real line into three intervals, namely $D_1 = [0, a_1)$, $D_2 = [a_1, a_2)$ and $D_3 = [a_2, \infty)$, and yields a sample of grouped and ungrouped data.

We again pretend that $X_j^{(\lambda_0)} \sim N(\mu_0, \sigma_0^2)$, $j = 1, \ldots, n$ for some parameter value $\underset{\sim}{\theta}_0 = (\mu_0, \sigma_0, \lambda_0)'$. The log-likelihood of the mixed sample of size n is then

$$\ell_n(\underset{\sim}{\theta}|\underset{\sim}{x}) = \log(n!) - \sum_{i=1}^{3} \log(n_i!) + n_1 \log[p_1(\underset{\sim}{\theta})] + n_3 \log[p_3(\underset{\sim}{\theta})] - \tfrac{1}{2} n_2 \log(2\pi) - n_2 \log(\sigma) - \tfrac{1}{2} \sum_{j=n_1+1}^{n-n_3} \left(\frac{x_j^{(\lambda)} - \mu}{\sigma} \right)^2 + (\lambda - 1) \sum_{j=n_1+1}^{n-n_3} \log(x_j).$$

with $n_1 + n_2 + n_3 = n$ and where

$$p_1(\underset{\sim}{\theta}) = \Phi\left(\frac{a_1^{(\lambda)} - \mu}{\sigma} \right) \text{ and } p_3(\underset{\sim}{\theta}) = 1 - \Phi\left(\frac{a_2^{(\lambda)} - \mu}{\sigma} \right). \qquad (6)$$

Let

$$\ell_{n_2}^{(u)}(\underset{\sim}{\theta}|x) = -\tfrac{1}{2} n_2 \log(2\pi) - n_2 \log(\sigma) - \tfrac{1}{2} \sum_{j=n_1+1}^{n-n_3} \left(\frac{x_j^{(\lambda)} - \mu}{\sigma} \right)^2 + (\lambda - 1) \sum_{j=n_1+1}^{n-n_3} \log(x_j)$$

be the part of the log-likelihood linked to the ungrouped data. Again consistency to the minimum information number solution and asymptotic normality follow.

Theorem 5. Let $q_i = \int_{D_i} g(x)dx$ for $i = 1,2,3$ and let $p_1(\underset{\sim}{\theta})$ and $p_3(\underset{\sim}{\theta})$ be as in (6). If

(i) the parameter space Ω is the compact set defined in Theorem 4,

(ii) the moments $E_g(X^{2a})$ and $E_g(X^{2b})$ are finite,

(iii) $H(\underset{\sim}{\theta}) + E_g[I_{D_2}(X)\ell_1^{(u)}(\underset{\sim}{\theta}|X)]$ has a unique global maximum at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$, where $H(\underset{\sim}{\theta}) = q_1 \log\left[\frac{p_1(\underset{\sim}{\theta})}{q_1}\right] + q_3 \log\left[\frac{p_3(\underset{\sim}{\theta})}{q_3}\right]$.

Then, (1) $\hat{\underset{\sim}{\theta}}_n \xrightarrow{a.s.} \underset{\sim}{\theta}_0$ as $n \to \infty$.

Furthermore, if:

(iv) $\underset{\sim}{\theta}_0$ is an interior point of Ω,

(v) both $E_g[X^a \log(X)]^2$ and $E_g[X^b \log(X)]^2$ are finite,

(vi) $\nabla H(\underset{\sim}{\theta}_0) + E_g[I_{D_2}(X) \nabla \ell_1^{(u)}(\underset{\sim}{\theta}_0|X)] = \underset{\sim}{0}$,

(vii) $V = (\nabla^2 H(\underset{\sim}{\theta}_0) + E_g[I_{D_2}(X)\nabla^2 \ell_1^{(u)}(\underset{\sim}{\theta}_0|X)])^{-1}$ exists.

Then, (2) $\sqrt{n}\,(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{d} N_3(\underset{\sim}{0}, VWV')$ as $n \to \infty$, where $W = (w_{uv})_{3\times 3}$ is defined by

$$w_{uv} = q_1 \left(\frac{\partial \log[p_1(\underset{\sim}{\theta})]}{\partial \theta_u}\bigg|_{\underset{\sim}{\theta}_0}\right)\left(\frac{\partial \log[p_1(\underset{\sim}{\theta})]}{\partial \theta_v}\bigg|_{\underset{\sim}{\theta}_0}\right)$$

$$+ q_3 \left(\frac{\partial \log[p_3(\underset{\sim}{\theta})]}{\partial \theta_u}\bigg|_{\underset{\sim}{\theta}_0}\right)\left(\frac{\partial \log[p_3(\underset{\sim}{\theta})]}{\partial \theta_v}\bigg|_{\underset{\sim}{\theta}_0}\right)$$

$$+ E_g\left[I_{D_2}(X)\left(\frac{\partial \ell_1^{(u)}(\underset{\sim}{\theta}|X)}{\partial \theta_u}\bigg|_{\underset{\sim}{\theta}_0}\right)\left(\frac{\partial \ell_1^{(u)}(\underset{\sim}{\theta}|X)}{\partial \theta_v}\bigg|_{\underset{\sim}{\theta}_0}\right)\right].$$

Although we have stated these results for Type I censoring, those for Type II are similar.

Figure 1

Histogram and Inverted pdf in the Original Scale

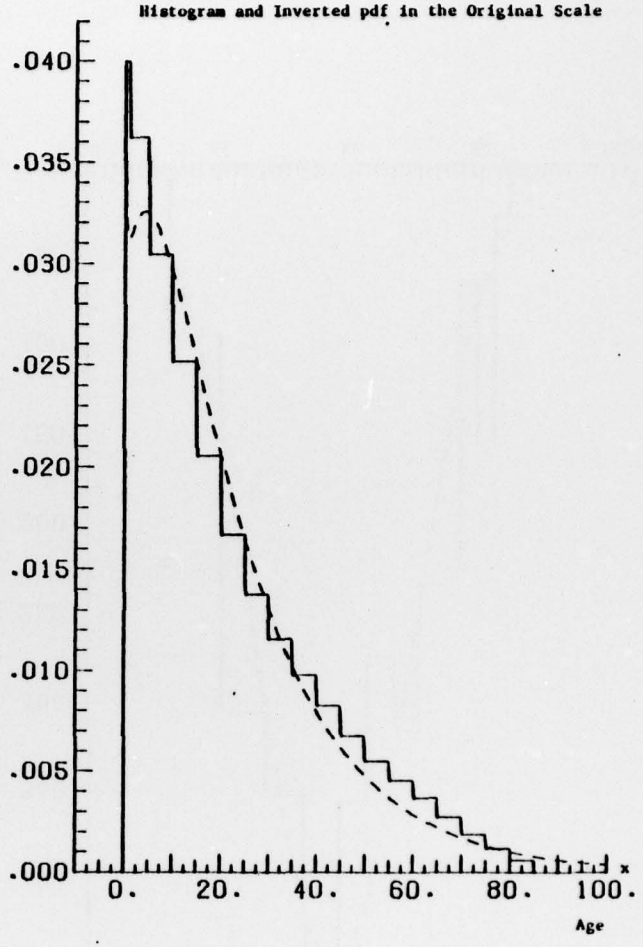


Figure 2

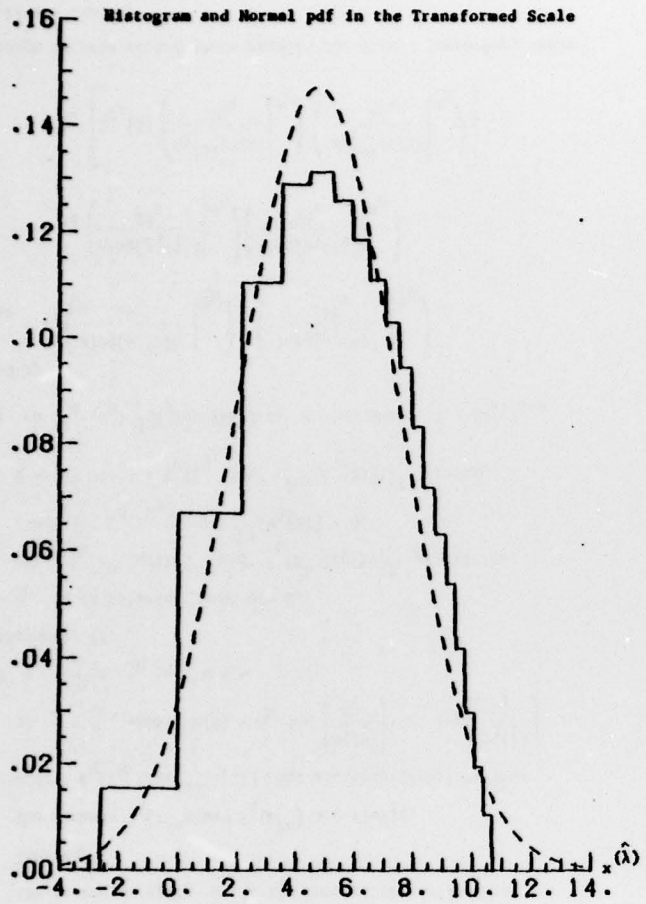


Figure 3

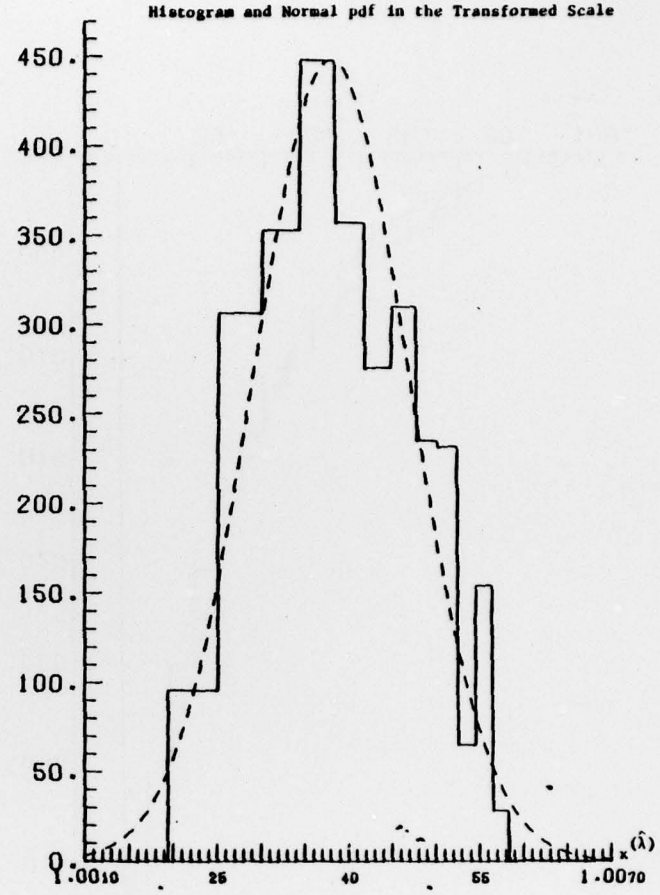
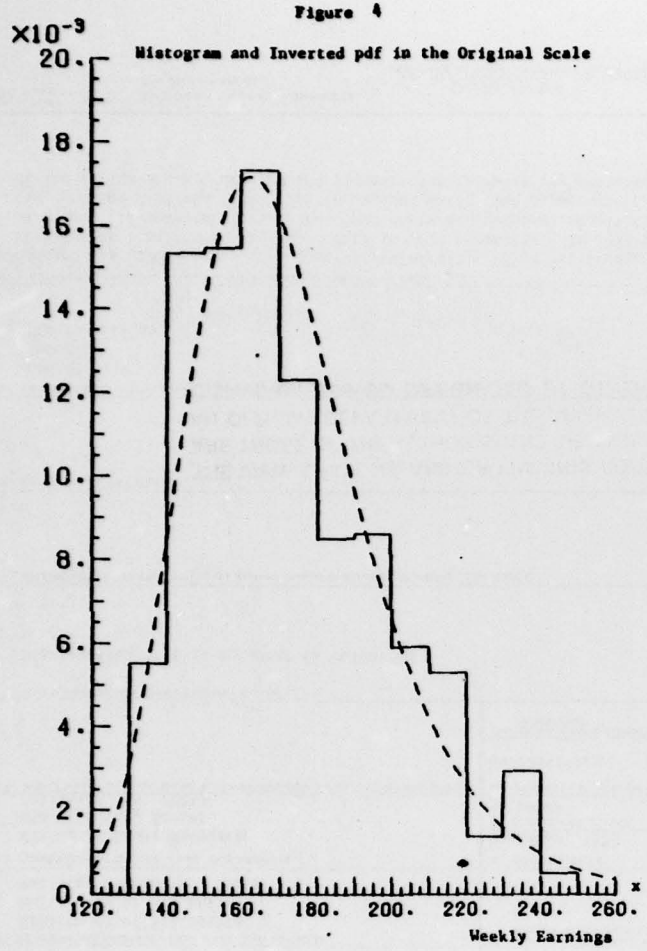


Figure 4

Histogram and Inverted pdf in the Original Scale



Bibliography

Benn, R.T. and Sidebottom, S. (1976). "Algorithm AS-95: Maximum likelihood estimation of location and scale parameters from grouped data." Appl. Statist. 25, 88-93.

Box, G.E.P. and Cox, D.R. (1964). "An analysis of transformations." J.R. Statist. Soc. B-26, 211-52.

Draper, N.R. and Cox, D.R. (1969). "On distributions and their transformation to normality." J. Royal Statist. Soc. B-31, 472-76.

Guerrero, G.V.M. (1979). "Extensions of the Box-Cox Transformation to Grouped-Data Situations." Ph.D. Thesis, University of Wisconsin-Madison.

Hernández, A.F. and Johnson, R.A. (1979). "Transformation of a discrete distribution to near normality." TR No. 546, Dept. of Statistics, U. of Wisconsin-Madison.

Kale, K.B. (1963). "Some remarks on a method of maximum-likelihood estimation proposed by Richards." J.Royal Statist. Soc. B-25, 209-12.

Keyfitz, N. and Flieger, W. (1971). Population, Facts and Methods of Demography. W.H. Freeman and Co.

Lindley, D.V. (1950). "Grouping corrections and maximum likelihood equations." Proc. Camb. Philos. Soc. 46, 106-10.

Richards, F.S.G. (1961). "A method of maximum likelihood estimation." J. Royal Statist. Soc. B-23, 469-75.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Technical Report No. 542	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): TRANSFORMATION OF GROUPED OR CENSORED DATA TO NEAR NORMALITY		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Victor M. Guerrero Richard A. Johnson		8. CONTRACT OR GRANT NUMBER(s): ONR Grant No. N00014-78-C-072 (Also funded by Army Res. Off
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics University of Wisconsin Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research 800 N. Quincy Street Arlington, VA 22217		12. REPORT DATE: November, 1979 13. NUMBER OF PAGES: 18 pages
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Grouped Data
Sensoring
Transformations
an analysis of transformations

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Box and Cox (1964) proposed a power transformation which has proven utility for tranforming ungrouped data to near normality. In this paper, we extend its applicability to two frequently encountered situations: (i) grouped data and (ii) Type I censored data. The asymptotic properties of the estimators are derived and illustrative examples are presented.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)